**InfVis — Platform-Independent Visual Data Mining of Multidimensional Chemical Data Sets**

Frank Oellien,* ‡ Wolf-Dietrich Ihlenfeldt, ‡ and Johann Gasteiger

Computer-Chemie-Centrum, University of Erlangen-Nuremberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany.

Received May 19, 2005

The tremendous increase of chemical data sets, both in size and number, and the simultaneous desire to speed up the drug discovery process has resulted in an increasing need for a new generation of computational tools that assist in the extraction of information from data and allow for rapid and in-depth data mining. During recent years, visual data mining has become an important tool within the life sciences and drug discovery area with the potential to help avoiding data analysis from turning into a bottleneck. In this paper, we present InfVis, a platform-independent visual data mining tool for chemists, who usually only have little experience with classical data mining tools, for the visualization, exploration, and analysis of multivariate data sets. InfVis represents multidimensional data sets by using intuitive 3D glyph information visualization techniques. Interactive and dynamic tools such as dynamic query devices allow real-time, interactive data set manipulations and support the user in the identification of relationships and patterns. InfVis has been implemented in Java and Java3D and can be run on a broad range of platforms and operating systems. It can also be embedded as an applet in Web-based interfaces. We will present in this paper examples detailing the analysis of a reaction database that demonstrate how InfVis assists chemists in identifying and extracting hidden information.

**INTRODUCTION**

More than in any other scientific discipline, the day-to-day routine in chemistry—and in particular in drug design—is based on the retrieval and analysis of existing information and knowledge. For example, medicinal chemists are using experience and available knowledge about structure—activity relationships, synthesis rules, solubility values, ADME/Tox properties, and further information during the R&D process of a new drug candidate. During this process, not only existing knowledge will be used but also, in addition, massive amounts of new data will be generated, as 15 years of development and over 800 million U.S. dollars have to be spent.¹ To reduce the time and costs expended on the drug discovery process and to increase the number of potential drugs, automated experimental designs and techniques, such as combinatorial chemistry, automated parallel synthesis, high-throughput screening, and genome analysis of protein and gene expression arrays have been developed. Today, these technologies are playing a central role in finding new leads in the drug discovery process. Pharmaceutical companies are now generating several magnitudes more data by applying the above methods than by using all other conventional approaches. Usually, this information is stored in databases and can be retrieved during the drug discovery process, when required. In light of the continuous increase in the size of these databases, an additional, much more important fact has been realized: data sets may contain hidden information that has not explicitly been entered. This information is hidden in relations between all data records, such as structural or biological similarities and structure—activity relationships. These patterns can be used as a key for generating new chemical models or for the prediction of biological activities or ADME properties. Therefore, the primary goals of the drug discovery process are pattern recognition and mining of hidden information, with the aim of gaining new chemical knowledge.² This process is well-known as Knowledge Discovery in Databases (KDD) or also Data Mining (DM).³

As a matter of fact, the data mining process is nothing new in the field of chemical research. Chemists have always analyzed their own data or information reported in the literature in order to find rules that allow them to build new models and make predictions. Initially, classical statistical methods were applied to obtain new knowledge from data sets. This work, which usually could only be done by experts, was a time-consuming process. Furthermore, the options of obtaining new knowledge were limited. Only the advent of high performance computer-supported systems made new and more advanced analysis approaches possible. These Machine Learning methods now allow for the handling of more complex data mining problems.

Because in many cases it is not possible to describe complex relations between data objects within chemical databases by just one or a few data dimensions, chemical data sets usually contain a large number of molecular descriptors. Therefore, the high dimensionality of data is another problem in addition to the database size during the chemical data mining process. Because of the high-

---

¹ Corresponding author phone: +49 (0)6130 948365; e-mail: Frank.Oellien@intervet.com.
² Present address: Xemistry GmbH, Auf den Stieden 8, D-35094 Lahntal, Germany.
³ Present address: Xemistry GmbH, Auf den Stieden 8, D-35094 Lahntal, Germany.
dimensional character, chemical data sets above all require specific multidimensional analysis methods. Today, chemoinformatics and molecular modeling scientists can choose from a broad range of multidimensional data mining tools, including hierarchical clustering, principal component analysis, multidimensional scaling, neural networks, genetic algorithms, support vector machines, and self-organizing feature maps.

Looking at the raw output of data mining tools, which usually consists of numerical and textual values in tabular form, it becomes clear that these abstract data representations are not immediately suitable for the interpretation and detection of patterns, relationships, or outliers. During the ages, man has always employed a different tool to understand such abstract data sets or similar complex situations—the eyes. Of all the sense organs, the human eye and the associated visual cortex display the largest bandwidth when gathering information. By visualizing the abstract, tabular data mining output, mentioned above, graphically in the form of a two-dimensional scatter-plot and encoding the data values by means of color, size, and xy-positions, it is possible to discover patterns and relations very quickly. This phenomenon is well expressed in the idiom ‘A picture paints a thousand words’. Therefore, methods for data mining have always been closely modeled on visualization approaches. While low-dimensional and small data sets such as principal component analysis results were visualized at rather an early stage by using two-dimensional plots, effective interactive visualization of large, multidimensional data sets has become possible only upon the advances in processor speed and hardware-accelerated graphic cards. The development of advanced methods and approaches for data visualization has evolved into a specific research field, called information visualization.

Information visualization techniques have not only been used for the graphical representation of classical data mining tool results. The human capability to visually recognize patterns and relations, which is still unsurpassed by computer methods, has also been used to establish an independent subfield of data mining—visual data mining. Visual data mining allows for the direct integration of the user and his capabilities into the analysis process, without using any other conventional data mining approaches. Because the application presented in this paper is based on a visual data mining approach, in the following subsection, we will present an overview of the fundamentals of this data mining technique.

**Visual Data Mining.** Although data mining applications have become more and more automated, it is not possible to effectively extract new knowledge from data without interaction with the user. Especially when looking at complex analysis questions, a solution can often only be found if human intuition, flexibility, creativity, and expert knowledge are integrated into the data mining process. Unfortunately, in many instances, conventional computer-supported data mining approaches are so-called ‘black-box’ systems, which only allow for limited or no interaction with the user. Furthermore, such applications often require a great deal of experience and good knowledge about the data mining tool itself, because of their expert and mathematical character. Therefore, the usage of such tools is often limited to specialists such as chemoinformaticians and statisticians. This situation can at times be unsatisfactory: medicinal chemists wanting to gain insight into their data sets need the assistance of their colleagues in the chemoinformatics department to analyze the data and obtain the required results. This process takes additional time and work. Visual data mining may be one solution to overcome this problem. Unlike classical data mining tools, visual data mining approaches can also be used by nonexperts, as visual exploration is a common human skill. With the aid of visual data mining, the analysis process becomes a dynamic process, which can be directly manipulated and influenced by the user. Furthermore, these techniques allow the analyst to gain a deeper and intuitive understanding which can lead to faster conclusions.

A different problem has also helped furthering the success of visual data mining tools in the life sciences: with the advent of new automated laboratory techniques and the resulting tremendous increase in the size and number of data sets, many conventional data mining approaches were unable to keep pace with the requirements. Therefore, new cost-effective computational tools for rapid data analysis are needed and may be found in the field of visual data mining.

The following subsection summarizes the potential advantages of visual data mining tools compared to classical data mining tools: (1) usage and incorporation of data-related and prior expert knowledge, (2) increased trust in the data mining process and retrieved results, (3) easy and intuitive data analysis process, and (4) handling of complex, problematic, or very large data sets, which cannot be analyzed with conventional data mining approaches.

Visual data mining tools are built using information visualization techniques. First approaches and basic rules for the use of such techniques for explorative data analysis have been introduced by Tufte and Bertin. Today, the user can choose from a wide range of different techniques and applications. However, in the life sciences in particular, approaches to visualize multidimensional data sets have met with considerable interest. The most important multivariate approaches and corresponding techniques are geometry-based approaches such as scatterplots, bar charts, and parallel coordinates, icon- and glyph-based techniques such as Chernoff faces and star glyphs, pixel- and voxel-oriented approaches such as recursive patterns and circle segments, and, finally, hierarchical and graph-based techniques such as dimensional stacking and cone tree visualizations. Furthermore, a multitude of hybrid approaches exists which represents combinations of the information technologies mentioned above. As one hybrid visualization approach, the 3D glyph technique, has been used as key technology within the application, presented in this paper, it should be discussed in detail.

The 3D glyph-based approach combines the 3D scatterplot technology and the glyph- or icon-based technology. Using this method, three data dimensions can be encoded using the three orthogonal axes. Similarly, all other data dimensions can be assigned to the graphical attributes of the glyphs such as shape, color, size, orientation, texture, or opacity. The mapping of data dimensions to retinal properties is called visual mapping. Figure 1 shows an overview of the most important retinal properties that can be used for this visual mapping procedure. In addition to the numbers of data dimensions that can be encoded using each of the retinal properties, the figure also contains examples for the mapping of continuous, numerical values and discrete, categorical data
analyzed by anybody interested. Pharmaceutical companies have started to use this technology to establish Intranet solutions such as data warehouses, chemical portals, or other client-server architectures. To provide medicinal chemists and other scientists in such companies with easy and standardized access to all available data sources, information management departments are currently investing considerable time and effort into data integration processes and the development of unified and in many cases Web-based, data retrieval, and analysis tools.

We have developed an application, InfVis, that allows for the platform-independent and Web-based visual data mining of multidimensional chemical data sets. Our intention was to provide an application that also enables laboratory-based chemists to explore, analyze, and mine the data sets generated by them. Using InfVis, chemists can visualize and interactively explore data sets in an intuitive manner. InfVis has been implemented in the platform-independent computer language Java and therefore can be integrated in existing Intranet solutions for data retrieval and analysis. Using the applet version, InfVis can also be implemented in global accessible Web services for data mining.

METHODS

Figure 2 shows a screenshot of the InfVis application. The graphical user interface can be divided into four sections—the menu bar (Figure 2, at the top), the visualization window (Figure 2, upper left), the tools window (Figure 2, upper right), and the mapping window (Figure 2, at the bottom). All windows are connected by so-called SplitPanes that allow for the resizing of each window. Therefore, any visualization window may be enlarged to the full screen, if desired.

The visual data mining tool has been implemented using the platform-independent programming language Java. Therefore, InfVis can be used as a standalone application on any platform or operating system that provides a Java Virtual Machine. Furthermore, InfVis can also be embedded as an applet into Web-based services. We employed the graphical Java library Swing, which is included in Java versions later than 1.2, to implement the user-friendly graphical user interface shown above. The rendering window, which contains the 3D glyph visualization, was programmed with the Java extension Java3D.

Although the complete tool consists of 106 Java classes with 39,000 lines of code, the application has only a size of 160 KB, allowing for fast load times even over the Internet or slow Intranet connections. To run the standalone application, a Java Virtual Machine that supports Java2 and the Java3D extension is required. These JVM engines are freely available. To run InfVis as a browser applet, a Java2 plug-in, which is supported by all modern browsers, is required. This plug-in can also be downloaded at no cost, if not already installed. However, many Internet browsers such as Netscape 6, Mozilla, or Firefox already contain a Java2 plug-in. In that case, the user only needs to install the additional Java3D extension, which unfortunately is not yet packaged by default in standard Web browser distributions.

Data Import and Management. The first step in the visual data mining pipeline is data import. InfVis contains two mechanisms for the import of data sets. First, the application provides access to commercial and free databases using the Java Database Connectivity interface (JDBC). In

Figures

Figure 1. Possible retinal properties available in the glyph-based information visualization technique.

Table

<table>
<thead>
<tr>
<th>Graphical Attribute</th>
<th>Dimensionality</th>
<th>Continuous data mapping</th>
<th>Discrete data mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colour</td>
<td>max. 3 dimensions (3 if colour opponent)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shape</td>
<td>max. 3 dimensions (x, y, z)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientation</td>
<td>3 dimensions (x, y, z)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texture</td>
<td>3 dimensions (contrast, size, orientation)</td>
<td>texture-morphing (unfavourable)</td>
<td></td>
</tr>
<tr>
<td>Animation</td>
<td>At least 2-3 dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinking</td>
<td>1 dimension</td>
<td>smooth blinking speed</td>
<td>blinking, non-blinking, defined steps</td>
</tr>
</tbody>
</table>
the current version, we only support connections to the MySQL database system. However, the functionality can easily be extended to other databases, embedding the corresponding database drivers into the application framework. Second, InfVis is capable of importing data sets from plain ASCII files that contain data values separated by semicolons. This format is supported as output format by many other applications such as Excel. In both cases, InfVis analyses the data set and determines meta information like the data type and labels/headers of the different data columns. In addition to the numerical and textual data values, InfVis is capable of identifying specialized data types such as hyperlinks or Base64-encoded images. These metadata are not accessible as data dimensions during the visual mining process, but InfVis will interpret and visualize this information in the detail tool described below.

An important feature of visual data mining tools is the management of different data sets. An exploration of the underlying data often requires data subsets that have been generated by subset selection, filter methods, or different data representation. Only by way of comparing such data sets does the analyst become capable of detecting relations and patterns in the data sets. In the InfVis application, not only raw data are stored within a data slot but also all user-defined settings such as graphical options and generated filters are kept with the data set.

Information Visualization. As already mentioned above, we have chosen the 3D glyph-based information visualization technique for our visual data mining approach. We firmly believe that this visualization technique has the best characteristics for application by users with little or no experience with data mining tools and visualization applications. As man has evolved to understand his environment as a multitude of systems that exist in three dimensions, three-dimensional visualizations such as the 3D glyph approach can be grasped easily and intuitively. Other more complex and abstract visualization methods such as parallel coordinates, require a more abstract way of thinking, may lead to visual confusion or refusal, and necessitate additional training. Furthermore, simple 2D or 3D scatterplots are familiar graph types. Therefore, 3D glyphs—a combination between 3D scatterplots and glyph-based techniques—combine high user acceptance with multidimensional data visualization capabilities. Figure 2 shows the InfVis application using a glyph-based data representation.

We have implemented the 3D glyph technique using SUNs Java3D API. As a high-level graphic interface, Java3D is based upon the low-level graphics standards OpenGL or DirectX. Therefore, Java3D can use the tremendous capabilities of current hardware-accelerated graphic cards which have been pushed by the demands of the gaming industry and are now available on most PCs. 3D visualizations by our tool even allow for the usage of virtual reality periphery and stereo glasses for further enhancement of viewing.

Besides the 3D glyph technology, InfVis also supports 3D bar charts and 3D scatterplots as additional alternative information visualization techniques. However, these techniques are less suitable for the visualization of multidimensional data sets than the 3D glyph approach. Since these are standard methods, we will not describe them further in this paper.

Visual Mapping. During data import, each column of the underlying data set is automatically mapped to a graphical attribute of the glyph or to one of the three orthogonal axes. InfVis provides six graphical attributes—three spatial axes as well as color, size, and shape of the 3D glyphs. As already mentioned above, this process is called visual mapping. During the import of a data set, the binding will automatically

![Figure 2. Screenshot of the InfVis application. The tool consists of the menu bar (top), the 3D render window (upper left), the tool area (upper right), and the control panel to change the visual mapping of data dimensions (bottom).](image-url)
follow a predefined order where each column will successively be mapped onto the ordered retinal attributes (x > y > z > size > color > shape). Of course, this will in many cases not lead to the optimal projection, because it is not possible for the application to predict the best way to map incoming multidimensional data to the attributes. The optimal projection largely depends on the nature of the underlying data, the way the visual data mining problem is formulated, and the preferences and visual capabilities of the analyst. Thus, the possibility to test and specify different graphical attributes for one data column is crucial for the production of a reasonable visualization. Furthermore, this functionality enables the user to visualize his data sets and the corresponding data dimensions in different ways, which is a fundamental requirement to explore unknown data sets.

Therefore, InfVis provides a general settings panel and one specific control panel for each graphical attribute to control the visual mapping process. Figure 3 shows a screenshot of the x-axis control panel, whereas the control panel for the glyph color is shown in the lower part of Figure 2. Every option panel contains a ‘Dimension Info’ section, where the user can retrieve fundamental information about the mapped data dimension, such as the number of single data points or the minimum and maximum values of the data column. The mapped data dimension can easily be changed by clicking on the data dimension name, thus opening a list with all selectable data dimensions.

Data variables can exist as discrete, categorical data types or as continuous numerical values. Typical categorical values are textual enumerations or a defined set of integers such as numerical categories and generally vary between one and several hundred data values per dimension. Data dimensions with numerical values, like float or decimal values, normally contain anywhere between 10 and an almost infinite number of records. InfVis automatically detects the type of the data variables and displays this information in the ‘Dimension Info’ section. Furthermore, each option panel contains a ‘Value Settings’ section, which informs about the data type. The visual mapping choices are restricted, depending on the data types. For example, if we want to map a set of four textual values such as benzene, amine, ethane, and propane to the color attribute of the glyphs, InfVis allows only the use of four categorical colors such as red, blue, yellow, and green. On the other hand, if we have a set of 200 melting points, the tool only supports the use of a continuous color scale ranging from one color (e.g. blue for low melting points) to another (e.g. red for high melting points). In borderline cases, the user may choose between the numerical or categorical representation of the data dimension, for example, in case there are 10 different numerical values.

Depending on the graphical attribute, the corresponding option panels contain several more options. On the axis option panels, the user can define thresholds, define the axis scaling, or manipulate the representation of the axis labels and tick marks (Figure 3). When mapping discrete data values onto the glyph color attribute, InfVis by default uses a set of standard colors (red, green, yellow, blue, black, white, pink, cyan, gray, brown, and magenta) per default, which have been optimized and are established in information visualization applications.13 Of course, these colors can also be changed by the user, if desired. In the case of numerical data values, InfVis uses continuous color gradients (Figure 2). The user can use predefined color scales such as the HUE model, spectral model, and the primary color model. Furthermore, he or she can also define his own two or three color gradients. In the glyph size panel, the user can define upper and lower thresholds for the glyph size. The visualization of data sets with more than six dimensions can also be achieved by combining the visual mapping technique with dynamic and interactive techniques, as described in the following section.

**Interactive and Dynamic Techniques.** InfVis contains four different kinds of interactive and dynamic techniques, which fully satisfy Shneiderman’s requirements: navigation tools, filter tools, selection tools, and detail tools.

With InfVis’s navigation tools, the user has two general options for navigating the data world. First, the analyst can use standard Java3D navigation functions such as zoom, rotation, and translation, which are controlled by the mouse. Second, InfVis provides an additional tool allowing the user to select standard data views. This tool is very useful, if the user has lost his/her bearings during the exploration process. Furthermore, it provides a fast switching method between different data views.

The most important interactive tools within InfVis are dynamic query devices,33,34 which are advanced filter tools that enable fast and easy data exploration. These tools are at the core of InfVis’s capability to explore data sets with more than six dimensions. A query device can be defined for each data dimension of the underlying data set and thus overcomes the limit of six dimensions. They can be understood as a graphical alternative to text-based SQL queries. However, in contrast to conventional SQL queries, such devices can be used in an intuitive manner and do not require specific database knowledge. By combining several dynamic query devices for different data dimensions, complex database queries can be represented.

Dynamic query devices can be controlled by different graphical user interfaces. InfVis provides the most common input devices such as item sliders, range sliders, checkboxes,
and radio buttons for the manipulation of these filters (Figure 4). The user can select one of these graphical devices and link it to a specific data dimension of his data set. Figure 4 shows the filter tool panel of InfVis with four user-defined dynamic query devices, which were mapped to different reaction conditions of a reaction database. The item slider on the top was added to filter the data set depending on the solvent used in the reaction, the range slider controls the display of data points belonging to a specific reaction yield range, the checkboxes control the display of data points depending on the checked resin types, and the radio buttons can be used to describe the influence of the reaction temperature. Although the user can link any graphical interface to any data dimension, it is useful to limit the use of check boxes and radio buttons to data dimensions with few data points, as one check box or radio button will be added for every single data value within this data dimension. Range sliders are recommended to filter data dimensions containing many data points. Item devices such as the item slider or the radio buttons, which allow for the selection of one specific data value from a data dimension, also provide an ‘All’ option (first tick of the item slider or additional radio button) to select all data values of a data dimension.

The interactive manipulation of these filter tools results in immediate and continuous updates of the data set in the visualization window. Only such data points of the whole data set, which fulfill the combined filter criteria of all dynamic query devices, are displayed in the 3D scene. By means of this real time interaction process, relations and patterns can be perused quickly.

In addition to the described filter tools, InfVis also includes general selection tools for the direct selection of data records by the user. These functions are especially useful, if identified patterns or data subsets are to be extracted as independent data sets. Selections can be performed in two ways. First, the analyst can define one or more three-dimensional selection boxes, which can be moved and resized within the 3D scene. All data points lying within these selection boxes will be copied to the subset. Using this technique, the user can select and combine specific local areas within his data set. Second, the user can define a new subset by choosing single data points in the 3D scene by using the mouse pointer. Depending on the visual mapping of the data set, different data points may have the same 3D coordinates. This situation might lead to problems using the second selection method, because the analyst might accidentally select multiple data points, when clicking on a specific 3D glyph. Other glyphs located in the same position will also be selected by the mouse pointer action. To address this problem, we have implemented an editable selection list. This list can be used by the user to finally control the selected data points and to remove data points selected at random. Finally, InfVis also allows for a combination of both selection methods.

In addition, InfVis provides a detail tool to display all information related to a specific data record (Figure 5). The user can select a specific data point within the 3D scene by using the mouse pointer. All textual and numerical values in the data set belonging to the selected data point will be presented. Furthermore, InfVis can also handle and display specific meta information types included in the imported data sets. Images such as 2D chemical structure plots, which have been embedded as Base64-coded strings into the data set, will be identified, interpreted, and visualized. Hyperlinks too will not be identified as simple text information but are accessible as graphical buttons, which will start a standard Internet browser with the corresponding hyperlink, if pressed. Among other things, this method allows for cross-references between a chemical structure within the imported data set and the same structure in a Web-accessible database.

DATA AND RESULTS

We have used two specific reaction data sets kindly provided to us by ChemCodes (now Nuada Pharmaceuticals) to show the capabilities of our application. Usually reaction databases have to be used with caution due to substantial differences in the quality of the reaction informa-
Figure 6. Aldol condensation of benzaldehyde and acetophenone.

A detailed description of ChemCodes’ proprietary synthesis design will not be presented here. All possible reaction combinations have been performed up to six times to ensure that the experimental values are highly consistent. From the resulting 480 single reactions (60 combinations * 6 repeats) ChemCodes provided us with a subset of 364 reactions, containing 63 out of the 80 possible different reaction combinations. The 364 reactions were combined to compute the means of the reaction yields, and outliers were removed during this process. The preprocessed data set with the resulting 63 single reactions was imported into InfVis.

Figure 7 shows a screenshot of the InfVis application after the import of the ChemCodes data set. We mapped the ChemCodes internal reaction category, which classifies the outcome of the performed reaction (Figure 7b), onto the x-axis, the solvents onto the y-axis, and the resins onto the z-axis. Furthermore, the size of the glyphs was used to represent the reaction yields, and the temperature was mapped onto the glyph shape. We did not use the remaining graphical attribute, the glyph color, to map another data dimension like the different reagents. In many cases, the use of all available graphical attributes to map different data dimensions is not required and sometimes even can lead to visual confusion. Instead, the glyph color was used simultaneously with the x-axis to represent the reaction category, which in our opinion yields a clearer representation. The two remaining data dimensions were linked to the visual mining by two dynamic query devices that are controlled via two item sliders (Figure 7a, right). Furthermore, an additional dynamic query device was created for the resins. Initially, all item sliders were in their initial position (‘All’ position) so that no data dimension was filtered and all data points were visible within the 3D scene (Figure 7a).

During the visual exploration of the data set, two general tendencies could easily be identified: reactions that have been performed at 60 °C and reactions on polystyrene showed significantly lower reaction yields (smaller glyphs in Figure 8) than reactions performed at 23 °C or reactions using Tentagel beads (larger glyphs in Figure 9). Although Figure 9 does not show exactly the complementary settings of the filter settings in Figure 8, the general temperature and resin dependencies are obvious by comparing these two images. The exact counterparts to the settings displayed in...
Figure 8 could be obtained within seconds by two mouse clicks: (a) changing the temperature item slider in Figure 8a from 60 °C to 25 °C and (b) changing the resin item slider in Figure 8b from PS (polystyrene) to TG (Tentagel).

In addition to the temperature-dependency of the reactions, the results also reveal a general problem inherent in the data set: some reactions were obviously not measured correctly, as is demonstrated by the small blue cubes at position 7 on the x-axis (Figure 8). The reason for this problem is that the data set was generated during the initial development of ChemCodes’ proprietary technology framework. During this time, the framework experienced some problems with the analysis of reactions performed on polystyrene, which later were overcome by changing analysis procedures.

Figure 9. InfVis application with filtered reaction data set: (a) 504 the best reaction yields could be obtained with KOH in ethanol/water (4:1) and ethanol on Tentagel resin (large red glyphs = high yields) and (b) reactions with LiOH on Tentagel show also good yields in ethanol but smaller yields with the ethanol/water solution. In DMSO or dioxane the reagents did not fully react.
To identify the most suitable activating reagent (base), we first filtered the data set to select only reaction conditions showing generally higher product yields (Tentagel and 25 °C). Subsequently, we examined the influence of each reagent by using the reagent item slider. A decrease in the reaction yields could be observed that followed the base order KOH > NaOMe > LiOH > no base > Pr2EtN. The results for two (KOH and LiOH) out of the six possible reagent slider settings (all bases, LiOH, KOH, NaOMe, Pr2EtN, and ‘no base’) are shown in Figure 9. The best reaction yields could be obtained with potassium hydroxide in ethanol or a mixture (4:1) of ethanol and water at 23 °C on Tentagel resin (Figure 9a). In DMSO, the yields were much smaller because of side reactions (small red glyphs or glyphs in category 1 indicate side reactions). When using dioxane as solvent, the reactions did not complete (green glyphs or glyphs in category 2). Compared to reactions with potassium hydroxide as base run in ethanol/water solvent (Figure 9a), lithium hydroxide-catalyzed reactions in the same solvent exhibit a clear increase of side products (smaller red glyph in Figure 9b). However, reactions run in pure ethanol also show high product amounts. In DMSO and dioxane, no significant yields could be obtained, due to incomplete reactions (Figure 9b, green spheres in category 2). An additional interesting difference between KOH- and LiOH-catalyzed reactions, which is not shown in the figures, could also be observed: in contrast to KOH catalyzed reactions, reactions using LiOH still generate high product yields and few side products at 60 °C. Selecting only reactions catalyzed by sodium methanolate, high yields in solutions with ethanol/water and DMSO could be observed. Corresponding reactions in ethanol and dioxane led to low product yields and a higher number of side reactions.

The results of these experiments gleaned from the visual exploration process are for the most part in agreement with the literature, general chemical reactivity tendencies, and rules such as the solvation effect of the different solvents. However, the data analyzed here provide reliable quantitative measures for these effects. On top of this, the visual data mining process presented here allows one to gain these insights and knowledge in a rapid and intuitive manner.

**Reaction Planning.** In addition to analyzing different reaction classes, ChemCodes’ reaction database was designed to solve the functional group compatibility problem. Knowledge of functional group compatibility makes selective reaction planning possible. For example, medicinal chemists often need to transform a specific functional group of a lead compound containing multiple other functional groups. The easiest way to find a suitable reaction is to access the reaction database and to retrieve specific reaction conditions that will allow only a selective reaction of the preferred functional group without interacting with any other functional groups present in the target compound. If such a process could be found, it would allow for a fast one-step reaction and would not require complex reaction sequences involving the attachment and removal of protecting groups.

ChemCodes provided us with a prototype data set to demonstrate the potential of such a database approach. The data set contained reaction data of 48 important functional groups that had been treated with 37 common activating/catalyzing reagents and 6 so-called quenchers in 11 solvents at 25 °C. The reagent library ranged from mild to strong agents to ensure a staggered activation of potentially reactive groups. Finally, the activated groups were transformed by the addition of the quenchers, which can be described as prototypical highly reactive counter-reagents.

We have used this experimental data set containing 126 115 reactions to identify reaction conditions that will allow for a reaction on the amino group of 3-aminobenzaldehyde, without interacting with the aldehyde group or the phenyl ring system of the compound (Figure 10). Furthermore, only such reaction conditions should be explored that will lead to a single main product and not to any additional side products. It must be mentioned that the provided data set was generated by very early experimental and analytical techniques during the development of the reaction mapping and analysis technology and that the results should not be applied directly to real reactions in the laboratory—there are obvious problems with the data as we received it.

First, we isolated a subset containing 7326 reactions that correspond to reactions at the amino group, aldehyde group, and phenyl ring system. The resulting data set was imported into InfVis. For the visual exploration, the number of products was mapped onto the x-axis, the solvents were mapped onto the y-axis, and the three functional groups were mapped onto the z-axis. Furthermore, the glyph shape was used to represent the product count, and the color was used in parallel with the z-axis to differentiate between the three possible reaction locations (amino group, aldehyde group, and phenyl ring system). The glyph size was not used in this study. All data points were visualized by glyphs with a uniform standard size. By reducing the number of used graphical attributes, the 3D scene became less convoluted. This example clearly shows that the visual mapping process strongly depends on the nature of the underlying data set and the data mining question to solve. The two remaining data dimensions, reagents and quenchers, were embedded by two dynamic query devices: the reagents could be changed by an item slider and the quenchers could be selected by checkboxes (Figure 11, right).

By using the dynamic filters, all possible reagents—quencher combinations were examined. For every combination we analyzed the 3D scene to identify solvents that lead to the desired result (only one product for the amino group and no reactions for the aldehyde group and the phenyl ring system). Figure 11 shows a screenshot of InfVis with a combination of reaction conditions which meet the requirements: using 1,3-disopropylcarbodiimide as activating reagent and 1-phenyl-2-thiourea as quencher, two solvents, toluene and N,N-dimethylformamide, could be identified where only the amino group will react. After the analysis of all possible combinations, 23 reaction conditions could be identified that lead only to reactions on the amino group.

The results of this study have to be interpreted with care. Inductive or electronic effects such as the substitution patterns (ortho, meta, para) could not be taken into account, because the data set did not provide for variations thereof. Full
production experiments as envisioned by ChemCodes would obviously include many more structural variations. Although the chemical reactivity of the functional groups has only been considered insufficiently and was highly generalized, the visual mining process shows the potential of ChemCodes’ reaction database.

DISCUSSION

With the tremendous increase in the number and size of chemical data sets and the simultaneous desire to speed up the drug discovery process there is a need for computational tools for rapid and in-depth data mining. During the last 5 years, visual data mining has become an important technique within the life sciences. In contrast to conventional data mining approaches, visual data mining tools allow nonexperts to explore and analyze complex data sets and to identify relations and patterns within those data sets. Several companies have identified the usefulness of visual data analysis tools within the life science sector and have developed a number of visual data mining tools that will be compared with the InfVis application in the following paragraphs. It must be noted that the development of InfVis commenced at a time when many of the advanced features of current commercial applications in the same area were not yet available, or even the applications themselves were not yet unveiled, and their future direction of development as well as their impact was not clear.

One of the first visual data mining applications that today has become the most successful visual mining tool in the drug discovery area is Spotfire.38–40 Even though Spotfire in its early versions had not been designed for the chemical industry, its high acceptance in the life science community led to the implementation of several specific modules for the use in lead optimization, lead discovery, and target discovery such as a structure viewer and functional genomics functions in recent years. One reason for the success of Spotfire was the use of intuitive dynamic query devices similar to those which were implemented in the InfVis application. Furthermore, Spotfire is one of the most powerful tools on the market. It provides a large number of different options and information visualization techniques such as scatterplots, parallel coordinates, and heat maps. However, its large functionality requires training courses and its potential to provide a set of different visualization displays has not always proven to be a suitable feature for inexperienced users.41 To solve a given task, beginners had to choose one specific visualization technique to find the right relations and patterns. However, it often takes a long time for nonexperts to identify the right visualization technique, and in some cases the users failed to change a wrong display technique even if they have arrived at a dead end.41 Today, most problems described above have been solved in current Spotfire versions. For example, Spotfire has implemented a tutorial framework into the application which can be set up by company-internal experts and which will lead beginners through their data mining task. Spotfire can be used as a stand-alone application and also as a client-server application on Microsoft Windows platforms. A platform-independent or Web-based deployment like in InfVis is not possible.

An application similar to Spotfire has been introduced by Partek.42 Partek not only uses an interactive spreadsheet for data representation but also provides a three-dimensional...
scatterplot display. As one of the first commercial applications, Partek has combined the visual data mining approach with classical data mining techniques. Besides statistical methods such as principle component analysis and multidimensional scaling, the analyst can also select a set of machine learning technique tools such as neural networks or genetic algorithms. Detail information can be shown in Microsoft’s Internet Explorer. The application also provides several database interfaces and tools for the embedding of chemical structures. Like Spotfire, Partek can only be run on Microsoft Windows machines. A client-server mode was not supported by Partek when we developed the InfVis application.

Miner3D is another application that uses 3D glyphs and dynamic filters to analyze data sets. During the development of InfVis, Miner3D was only available as Miner3D for Excel. This version had severe data import limitations, because the analyst could only import data sets by using Miner3D’s own data format or Excel tables. Like InfVis, Miner3D primarily not only uses the 3D glyph technology to visualize the underlying data but also provides bar charts as alternative representations. Compared to InfVis, Miner3D supports many retinal properties such as textures and transparency in the visual mapping process. It allows for the mapping of sounds and speech to the data dimensions. The analysis tool makes use of the 3D hardware capabilities of current computers just like like InfVis. Miner3D supported a plug-in for interactive visualization within MS Internet Explorer. Since InfVis development was finished, new versions of Miner3D like Miner3D Enterprise have been published. They now also support data picking and access to databases.

The most important difference and advantage of InfVis is, compared to the applications described above, the platform-independent architecture of the visual data mining tool. The application can be run as a standalone application on several platforms and operating systems and can also be used as an applet in Web applications. Therefore, InfVis can easily be integrated into existing Web-based data retrieval and data warehousing applications. InfVis was tested successfully on several Microsoft Windows operating system releases, on SGI/IRIX-based platforms, and on Linux (Suse 9.0 with Java 1.3.1 from Blackdown). Furthermore, the application has been embedded as an applet into a noncommercial Web application for mining of the NCI antitumor database. To our knowledge, InfVis is the first visual data mining application that can be used in a platform-independent way and that uses 3D hardware capabilities of current graphic hardware. Furthermore, InfVis was one of the first chemical applications leveraging the Java3D standard.

From the very beginning, InfVis was meant to be used by chemists—who usually are not experts in chemoinformatics or statistics tools—without any requirement of time-consuming training. Therefore, we tried to keep things simple during the development of the application, without losing the power of the visual data mining approach. The most important issue in solving this problem was the combination of the intuitive 3D glyph visualization technology with the power of dynamic query devices. Furthermore, additional features such as embedded statistical methods or clustering algorithms were not implemented into InfVis. One critical point was the selection of a reasonable information visualization technique. Of course, all these techniques are based on the natural pattern recognition capability of the human visual cortex and therefore are easy to understand; however, we decided to use the 3D glyph technology. In our opinion, this technique is the most intuitive approach for inexperienced users, because it could be shown that the scatterplot technique, which is similar to 3D glyphs, is the most suitable tool for beginners to identify patterns and relations in data sets. Furthermore, we have limited the number of graphical attributes that can be used by our application. InfVis provides six graphical attributes—three spatial axes and color, size, and shape of the 3D glyphs. Of course it is possible to map up to eight graphical attributes to a glyph, but too many graphical attributes can lead to overcrowded data displays, which finally may confuse and overwhelm inexperienced users in particular. Furthermore, in many cases, using as many graphical attributes as possible is not the best way to visualize multidimensional data sets and even can lead to a more difficult exploration process.

InfVis has not been designed as a tool only for a specific field in the area of chemistry, such as e.g. medicinal chemistry. It is intended to be used by all chemists who want to analyze multidimensional data sets. Because of its open character, InfVis can also be used by analysts from other disciplines.

CONCLUSION

The analysis of the vast amount of data generated by current automated laboratory techniques has become a critical bottleneck in chemistry today. Visual data mining approaches seem to be appropriate techniques to face this problem. In this article, we have presented a Java-based visual data mining tool. It allows for an easy and intuitive exploration and analysis of multidimensional data sets. Traditionally, these data mining tasks were performed by experts such as chemoinformaticians with the aid of classical data mining tools. With InfVis, chemists can explore and analyze many of their data sets on their own. This is an efficient approach. The specific knowledge of the creator of the data set can be leveraged directly. The incorporation of expert knowledge about the data set assists in the efficient evaluation of the results in the context of the respective chemical question. Ultimately, there is reason to believe that this yields a higher quality of the recognized patterns and relations and results in increased trust in the data mining process and the retrieved results.

By providing a 3D glyph visualization technique, we implemented a method that is well-known and accepted by chemists, because of its intuitively understood relationship to scatterplot graphs. Our experience shows that these techniques seem to be a good solution to allow beginners and nonexperts to visually explore data sets. InfVis also implements a dynamic query device technique. Combining this technique with the 3D glyph approach, we were able to develop a powerful visual exploration tool that gives the user maximum control of the data mining process. We found that dynamic query filters are efficient means to control the number of graphical attributes and data points in the 3D display, decreasing the problem of visual confusion, without losing the capability of exploring multidimensional data sets. This combination not only provides beginners with an easy tool for data analysis but also gives experts a fast tool to
explore more complex data sets. We have shown, using practical examples from high-throughput reaction screening, that interesting information can be extracted quickly from complex and large data sets by performing a few straightforward InfVis operations.

Because of its platform-independent architecture, InfVis can be run on any computer system with a Java Virtual Machine. Furthermore, InfVis can be embedded as an applet within Web-based applications. InfVis can be freely accessed on the Web as a component of an Internet accessible computational tool (http://cactus.nci.nih.gov/3DMiner) for the analysis of structure—activity relations of the NCI antitumor database.

ACKNOWLEDGMENT

The authors wish to thank Nuada Pharmaceuticals (formerly ChemCodes Inc.) for the reaction data sets discussed in the Experimental Section and the German Research Foundation (DFG) for financial support of this work that was funded part of the ChemVis project.44 ChemVis was a research project in the strategic research initiative “Distributed Processing and Exchange of Digital Documents” of the German Research Foundation. We also want to thank Prof. Thomas Ertl, Klaus Engel, and Guido Reina from the Visualization and Interactive Systems Group, Institute of Computer Science of the University of Stuttgart for their remarks and suggestions during the implementation of the InfVis application.

REFERENCES AND NOTES

(26) Kraus, M.; Ertl, T. Interactive data exploration with customized glyphs; In Proceedings of WSCG ’01; Přerov, Czech Republic, 2001; pp P20–P23.
(31) SUN, http://java.sun.com
(35) http://www.chemcodes.com (www.nuadapharm.com)
(36) Clausen, L.; Claparède, A. 1881, 14, 2463.
(37) Schmidt, J. G. Ber. 1881, 14, 1459.
(42) http://www.partek.com
(43) http://www.miner3d.com
(44) http://www2.chemie.uni-erlangen.de/projects/ChemVis C1050202K